Supplemental Materials

This supplementary material provides additional watermark visualizations for different models as referenced in Section Section IV (Experimental Results and Discussions) of Chapter D (Robustness Experiments) of the main manuscript. These results further demonstrate the robustness on the model of Simswap, FaceShifter, and MobileFSGAN. The watermark images presented here were generated using the same parameters as described in the main manuscript. The models are evaluated under the same conditions to ensure consistency in the results.

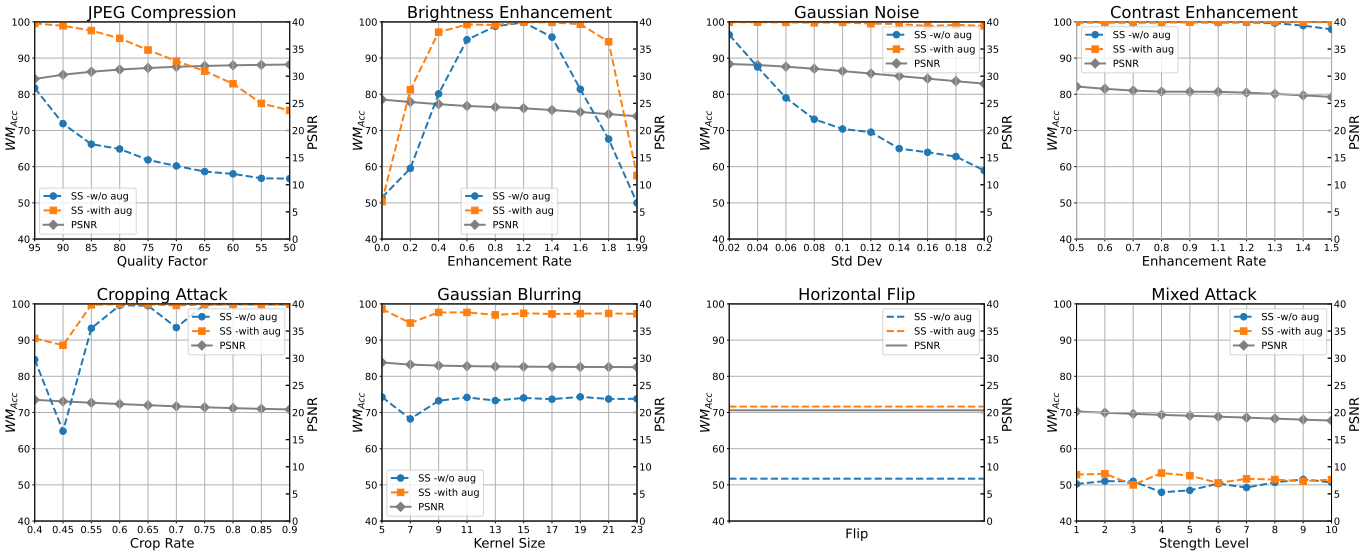

Fig. 1: Robustness of original and augmented Secure Swap SwimSwap model against different strengths of eight different image processing attacks. The grey line indicates average PSNR of output images after the attack.

As shown in Fig. 1, robustness of the SwinSwap model under eight image attacks is compared between the original SecureSwap model (SS - w/o aug) and the augmented SecureSwap model (SS - with aug). The augmented model consistently outperforms the original model in *AccWM*, particularly in the JPEG Compression and Brightness Enhancement attacks, where its accuracy remains high as attack intensity increases, while the original model's accuracy drops sharply. Under Gaussian Noise and Cropping attacks, the augmented model shows a slower decline. Across all attacks, the augmented model maintains better *AccWM* and PSNR values, demonstrating enhanced robustness.

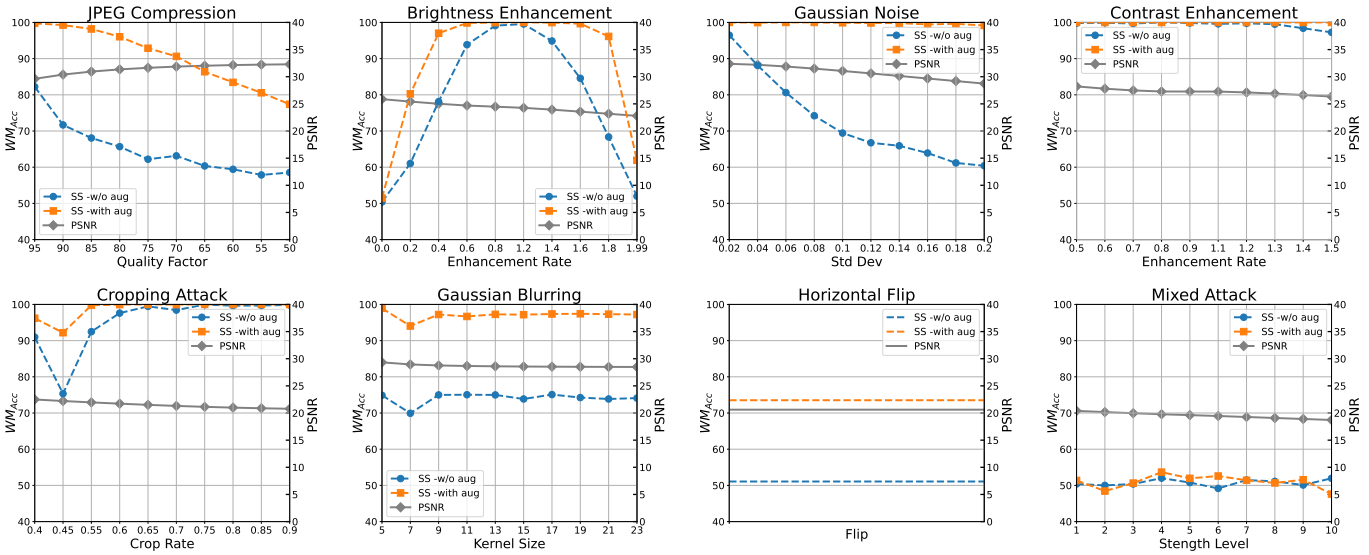

Fig. 2: Robustness of original and augmented Secure Swap FaceShifter model against different strengths of eight different image processing attacks. The grey line indicates average PSNR of output images after the attack.

As shown in Fig. 2, robustness of the FaceShifter model shows a similar pattern. The augmented model performs significantly better than the original model, especially in JPEG Compression and Brightness Enhancement, with higher Acc_{WM} as attack strength increases. The augmented model also exhibits a more gradual decline in accuracy under Gaussian Noise and Cropping attacks, while maintaining higher PSNR values across all other attacks.

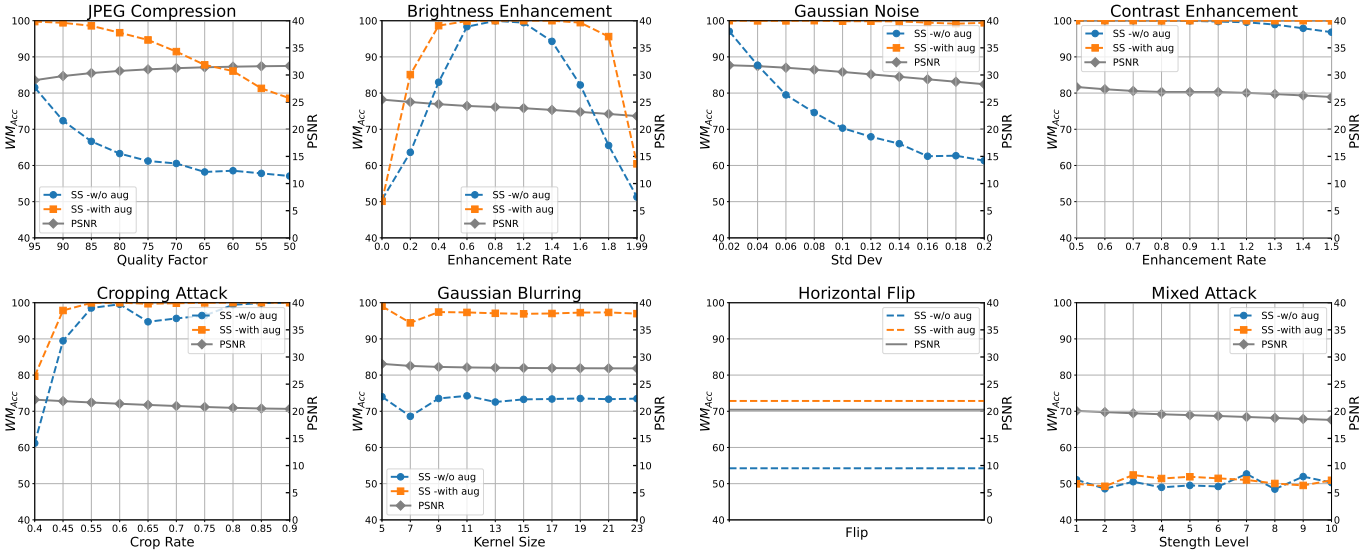

Fig. 3: Robustness of original and augmented Secure Swap MobileFSGAN model against different strengths of eight different image processing attacks. The grey line indicates average PSNR of output images after the attack.

As shown in Fig. 3, MobileFSGAN model results further confirm the advantages of the augmented model. It consistently achieves higher Acc_{WM} and PSNR values across all attack types, particularly in JPEG Compression, Brightness Enhancement, and Gaussian Noise attacks, where the original model's accuracy drops quickly. The augmented model's performance under all attacks demonstrates its superior robustness

Across all three models (SwinSwap, FaceShifter, and MobileFSGAN), the augmented SecureSwap models consistently demonstrate superior robustness against various image processing attacks compared to the non-augmented models. The augmented models maintain higher watermark extraction accuracy and PSNR across all attack types and intensities, indicating that the proposed augmentation method effectively enhances the robustness of the SecureSwap mechanism.